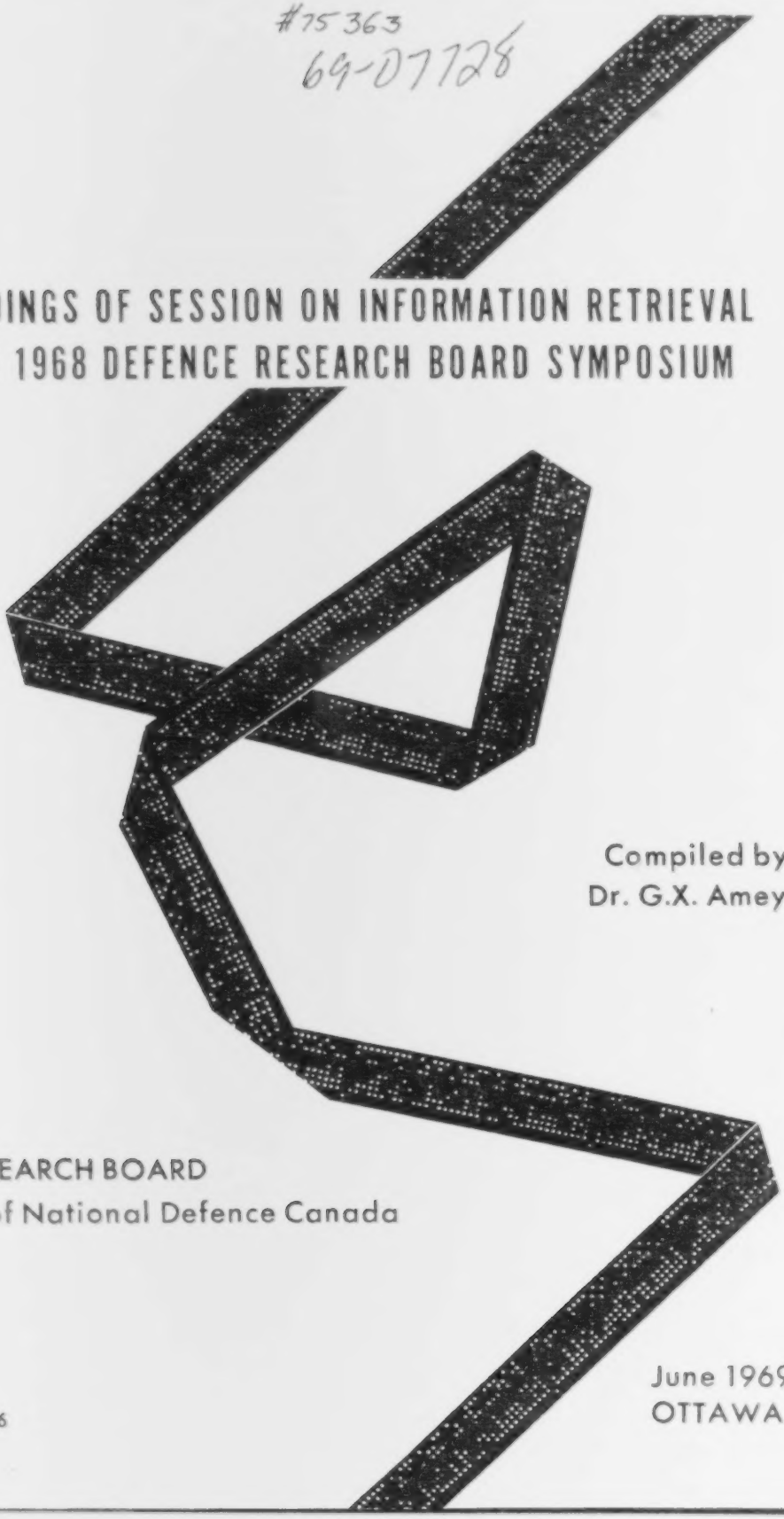
#75363
69-07128

PROCEEDINGS OF SESSION ON INFORMATION RETRIEVAL AT THE 1968 DEFENCE RESEARCH BOARD SYMPOSIUM



Compiled by
Dr. G.X. Amey

DEFENCE RESEARCH BOARD
Department of National Defence Canada

REPORT NO. DR 196

June 1969
OTTAWA

DEFENCE RESEARCH BOARD

DEPARTMENT OF NATIONAL DEFENCE
CANADA

PROCEEDINGS OF SESSION ON INFORMATION RETRIEVAL AT THE 1968 DEFENCE RESEARCH BOARD SYMPOSIUM

Compiled by

Dr. G.X. Amey

Report No. DR 196

Published June 1969
Reprinted August 1969
OTTAWA

69-011

105-03835/

TABLE OF CONTENTS

Foreword v

Computer Incompatibilities – A Software Solution 1

A Multiplicity of Computers 1

Incompatibilities 2

Example 2

Translation and Context 3

Summary 4

Information Retrieval for the Scientist – A DRES Experiment 5

Introduction 5

Experiences in Information Retrieval Using Chemical Abstracts 13

Magnetic Tapes 13

SOCRATES – An Information System for the Canadian Defence Community 18

References 25

File Design for Project SOCRATES 26

Introduction 26

Coding of Documents 27

The Computer Software Problem 28

The File Design 29

Key Word Selection 31

Future Systems 33

Communication Facilities for an On-line Information Retrieval System 36

Introduction 36

SOCRATES 36

Implementation 36

Implementation Phases 37

Communication Traffic 38

Terminal Characteristics 38

Communication Channel Speeds 39

SOCRATES Initial Communications Network 39

Operating Principle 40

Annual Communication Costs 43

User Benefits 44

The Future 44

Conclusion 46

FOREWORD

by

A.C. Jones

The information explosion has stimulated intense activity in development of new concepts, techniques and equipment for information processing. Information science is becoming recognised as a discipline. Information processing technology is developing rapidly. Practically every one is affected by the problem, and by the many solutions or expedients that are being developed. Mechanization and computer processing of document information is here to stay.

Defence scientific and technical information is a special case. Documents containing it can be obtained only under special agreements, used by selected persons for defence purposes only, and safeguarded at all times.

Much of what is being done in processing the open literature can be applied to defence information, but the complications introduced by these restrictions leave a great deal of development to be done by defence scientific information agencies.

The Defence Research Board has been conducting development work in this field. Some of it has been carried on by research scientists in the Board's laboratories, and some by information scientists in the Defence Scientific Information Service.

At the Board's 20th Annual Symposium, in Ottawa, the session held on the afternoon of 21 November, 1968 was devoted to six papers based on the Board's information work. These are reproduced in this report.

COMPUTER INCOMPATIBILITIES – A SOFTWARE SOLUTION

by

R.F. Rodgers

Defence Research Board
Ottawa, Ontario

(ABSTRACT)

Computer users requiring the use of more than one machine in solving an application run the risk of incompatibility between computers. This applies not only to computers produced by different manufacturers, but may be true of different machines produced by the same manufacturer. In applications involving the IBM 1401 standard character set, it was found that applications involving the CDC 3205, UNIVAC 1108, IBM 360 (of course) and FRIDEN 2301 flexowriter could produce, in the computer, not only different internal representation in every case, but even different printout in every case.

Higher level languages (fortran and programming language 1) were used by Defence Research Board to produce software conversions which either translated characters in strings or took certain specific actions upon reaching given characters. The process is illustrated by a simple edit. Widespread use of these techniques could open applications now considered infeasible.

A MULTIPLICITY OF COMPUTERS

Many computer applications involve the use of a single computer only. Other applications may not be so simple computerwise – and several machines, made by either one manufacturer or by several, are required. This situation could arise from the receipt of input from an independent supplier. It could also arise in-house from computer input/output limitations. In either case, the problem of compatibility rears its head.

An example of the second problem occurs in the Defence Research Board Directorate of Scientific Information Services. Information is key punched on the Friden 2301 flexowriter to produce 8-channel odd-parity punch paper tape. This tape is read by the Control Data Corporation 3205 computer to produce 7-channel even-parity Binary Coded Decimal magnetic tape card images. These tapes may be read either by the Univac 1108 computer to produce edited tape output, or by the International Business Machines 360 to produce 9-channel Extended Binary Coded Decimal Interchange Code print format records for program-control TN print chain listing.

Since the output of one computer is the input to another, we are left with the conversion problem. This problem is by no means one that was created by the IBM 360, but exists even among the so-called "1401 compatible" series of computers.

INCOMPATIBILITIES

To illustrate this point, let us consider the 12-7-8 triple punch on the standard 80-column punch card. The representations on the above-mentioned computers are given in the first slide, viz:

PUNCH CARD 12-7-8 CHARACTER

Computer	Internal Representation	Print Character
Even-parity 1401 tape	BA8421(77 octal)	—
IBM 360	4F hexadecimal	inclusive or
CDC 3205	37 octal	exclamation
Univac 1108	03 octal	numeration
IBM 1401	CBA8421	group mark

With a mixed-up situation like this, there is only one direction things can go, and that is the direction of improvement. However, until standards are imposed upon manufacturers, we must find a way around the incompatibilities.

EXAMPLE

An illustration may be made from the simplest of a number of programs written by Computation and Analysis Section at Defence Research Board. This program, like the others, is written in a higher-level language, viz., FORTRAN, and an excerpt from it appears in the next two slides, viz.,

EXCERPTS FROM PROGRAM RFR065, SLIDE 1 OF 2

```
C               PRØCESSING ØF CØLUMNS 1 TO 75 INCLUSIVE
                DØ5340WØRD=1,25
                D05330CHAR=0,12,6
                CALLGETIND(INCØNT,WØRD,CHAR,STATUS,INDEX)
                GØTØ(5080,5330),SWITCH
     5080       GØTØ(5330,5330,5310,5110,5280,5180,5180),INDEX
C               LETTER L
     5110       IF(STATUS.ØR.ALPHAB)
              A GØTØ5280
                STØRE=STØRE+1
                TCARD(STØRE)=2
                TWØRD(STØRE)=WØRD
                TCHAR(STØRE)=CHAR
                GØTØ5310
```

In this slide, all quantities are either fixed-point variables or logical variables. The GETIND is a subroutine which determines the input character under review, makes a substitution in those cases where that character is erroneous, sets the logical variable STATUS if so indicated, and always sets the variable INDEX to indicate the class of character. Two cases of the index are considered here: the letter L (INDEX=4) and the delimiter (INDEX=7).

For the letter L case, no further study is made if the letter is majuscule (STATUS=.TRUE.) or the character string containing L is non-numeric (ALPHAB=.TRUE.). If further action is taken, tables are filled with information regarding the existence and location of this minuscule L; such information will be used later when the next delimiter is found.

The action taken when such a delimiter appears is shown on the next slide:

EXCERPTS FROM PROGRAM RFR065, SLIDE 2 of 2

```
C                    DELIMITER
     5180            IF((.NØT.LENGTH).ØR.ALPHAB)
           A         GØTØ5230
                     IF(STØRE)5 230,5230,5210
     5210            DØ5220I=1, STØRE
                     TC=TCARD(I)
                     TW=TWØRD(I)
     5220            INCØ NT(TC,TW)=CØ NTNT(TCARD(I),TWØ RD(I)) +
           A         32*64**(5–TCHAR(I)/6)
     5230            LENGTH=.FALSE.
                     STØRE=0
                     ALPHAB=.FALSE.
                     GØTØ5330
```

For each delimiter (and the end of the record, per se, is not a delimiter), no further study is made if the field is of zero length (LENGTH=.FALSE.), i.e., the preceding character was also a delimiter, or if the string is non-numeric (ALPHAB=.TRUE.). If further action is taken, and tables indicate the presence of at least one minuscule L in the string, then that L is changed to 1 and the variables initialized for the next string.

TRANSLATION AND CONTEXT

These slides illustrate one aspect of character handling in this simple program. Two aspects will be noted. One aspect is character translation. The other aspect is context. Some pitfalls of each may be of interest.

Regarding translation, I had occasion recently to look up the word "anaptyxis" in the dictionary. This dictionary gave one, and only one definition, as shown on the next slide:

ANAPTYXIS≡

VOWEL EPENTHESIS

If you happen to know what vowel epenthesis is, you have just scored a triumph with this dictionary. If you don't know, you'll have to wait until the question period to ask. The important point to realize is that a translation must be meaningful to be significant.

Regarding context, there is the story of a man in deep personal trouble who decided to turn to the New Testament for guidance. Now this in itself is fine, but his method left something to be desired. He decided to open his book at random, and write down the first verse that caught his eye. After doing this three times, he would examine the verses written, and meticulously follow the advice set out before him.

On the first try, he opened his testament at Matthew, and read of the last act of Judas Iscariot. On the second try, he opened at Luke, and read of the conclusion of the story of the "Good Samaritan" told to the lawyer. On the third try, he opened at John and read of Jesus' admonition at the "Last Supper". The combined advice is shown in the next slide:

AND HE WENT OUT AND HANGED HIMSELF (Mt. XXVII 5)

GO AND DO LIKEWISE (Lu. X 37)

WHAT YOU ARE GOING TO DO, DO QUICKLY (Jn. XII 27)

Again, the importance of context requires a necessary knowledge of what precedes and what succeeds.

SUMMARY

Ideally, one would wish to have a standard character set employed by all computer manufacturers. At present, each new computer that has a different character set which renders a competitor's product obsolete gives that new competitor's manufacturer an edge in a competitive market. So we can't expect this problem to go away overnight. Instead, if we want a contemporary solution, we will have to do it ourselves either by hardware or software.

The above examples will, I hope, point out the way for software conversion while avoiding some of the pitfalls inherent in the subject. Thank you.

WARNING
The use of this information is permitted subject to recognition of proprietary and patent rights.

INFORMATION RETRIEVAL FOR THE SCIENTIST A DRES EXPERIMENT

by

D.J. Currie

INTRODUCTION

Bilbo Baggins said "Never laugh at a live dragon". For those of you who are not familiar with the writings of J.R.R. Tolkien, Bilbo Baggins is a hobbit, who with a company of dwarves goes off in search of a treasure stored in a mountain and guarded by Smaug, a live dragon. To some extent the classified literature is a treasure guarded by a dragon, a dragon who must be coped with before the wealth in the mountain can be obtained. But perhaps I can be more explicit. If, as happened to be the case a couple of years ago, a chemist, recently employed by the Board was asked to initiate research in a certain field, there was no way in which he could find out what classified information had already been published on that subject. To be sure DSIS has some broad keywords that could simplify the search, but some are very general (for example "Organic Compounds" might produce many hundreds of documents). The local library staff at DRES might happen to have some more specific information in their independent card file and the head of the department is likely to recall having read a key document on the subject, one with a yellow cover published a couple of years ago. But this is not really very helpful! The problem becomes more acute when, after this chemist has been with the Board some four or five years he asks whether the n.m.r. spectra of ethyl 2-chloro benzal acetoacetate has ever been measured. Perhaps he could have asked the n.m.r. specialist a few months earlier, but the specialist recently left the Board's employ. Maybe this specialist even left his card system but the key to it is known only to him.

Suffice it is to say that I, and most of my colleagues at Suffield recognized a need to develop a common information retrieval system for a segment of the classified literature – and at least find out what sort of dragon we must deal with in order to search the classified literature in depth, and retrospectively.

All scientists are being supplied with valuable information necessary for their jobs. Normally, a scientist will consistently remember some parts of this information very well: other parts of this information will be remembered poorly. Unless we can find a way of coding material that we normally recall badly, in terms of what we remember, our practical efficiency will be decreased.

Many attempts have been made to invent generalized codes to improve the capacity to remember, but for most of us some crutch, specifically an information transfer system, is required. By this I do not mean an information retrieval system in the sense of a broad base system designed to provide the scientist with current and

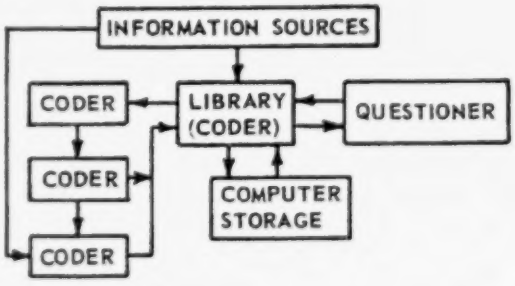
relevant information with which he is probably unfamiliar: I mean instead a system whereby a scientist in a defence laboratory can recall information which he himself has gleaned from many and varied sources:- classified documents, journals, and correspondence, to name just a few. Memory systems of this kind in use today include such things as personal card systems, and personal secretaries. Since most of us are not fortunate enough to have personal secretaries and in addition secretaries have not been computerized. . . yet, I will comment on card systems and their modification for computer storage and processing.

Most card systems are used to code facts or concepts to be recalled at some later date. Card systems have been found to be very useful in limited fields, especially if the coding has been formulated and is controlled by the person who is always going to use the system. In completely generalized and impersonal ways, however, card systems are apt to break down, largely because of their inevitable complexity and ambiguity.

This afternoon I want to discuss, in general terms, a computer-oriented card system that is designed to be used by many scientists in several diverse fields. Such a system for a small computer has been in use at Suffield since January of this year and, while still in the experimental stage, would appear to have many advantages both to the scientist and to the librarian.

Libraries, at least conventional libraries, alphabetically store cross-referenced concepts on many cards, while the scientist's punch card system is set up in terms of several concepts on a single card. For computer application the latter approach seemed more practical for storage. Moreover, numerically-coded concepts, like the holes in a card of a punch card system, are rapidly manipulated by computers, whereas alphabetic entries are not. In adopting numeric concept coding we avoided the need for checking extensive concept authority lists, the problems of relating broad and narrow terms and not least of all mis-spelled words.

The Information Retrieval System at DRES is concerned at present mainly with classified documents in the CW/BW field, but I hope this presentation will prove general enough that you can see the fundamental usefulness of this system.



SLIDE 1

Since January of this year, incoming documents to DRES are handled as follows:

(i) The document is received, accessioned, and recorded.

(ii) The Library staff enters certain housekeeping information about the document on a coding form which will be described in detail later.

(iii) The document and this coding form are circulated to all persons having an interest in the subject content of the document.

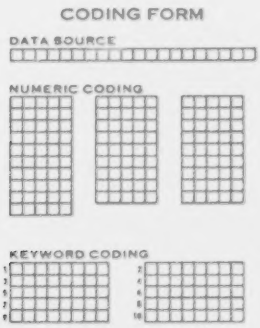
(iv) While reading the document, these persons may make entries on the form to code information of interest. In this way, the library staff uses persons who are perhaps better able to select pertinent information for storage. However, because of the limited field of interest of each coder, comprehensive coding may not be accomplished for the majority of documents. The assumption here is that any information that is coded is useful.

(v) The document and the form are passed back to the library with any recommendations for further circulation.

(vi) After all the interested persons have seen the document, the Information Services Group is responsible for keypunching the coded information and for disposition of the document.

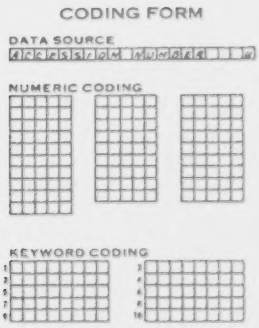
(vii) The information is stored on a disc in the computer.

The Document Coding Form that accompanies the document on its travels is presented in the next slide.



SLIDE 2

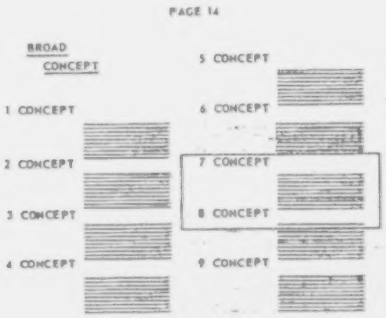
It consists of three parts: a location for data source, a location for numeric coding and a location for alphameric keyword coding.



SLIDE 3

The data source is labeled in the first section, and in Suffield this is generally the accession number of a classified document. A maximum of twenty alphameric characters are allowed and therefore journal references, correspondence, files, slides and most other sources can be entered. This information is not searchable and serves only as the address to be used in locating the source of information. In our system the twentieth character is used for the security classification alert.

As indicated, the document coding form allows for 28 numerically-coded concepts for each document. A concept dictionary was defined so that one narrow concept could be coded by five digits. The fourth slide represents a stylized version of a page of the concept dictionary.



SLIDE 4

This dictionary by the way is limited to 326 pages and at present 24 are being used. Each page is numbered and titled to indicate a very broad concept much like the Subject Field and Group Structure Index of Documents Digest. The pages are then divided into 9 sections, which are labeled to indicate a maximum of 9 different and numbered concepts relating to the broad concept defined by the page title. Then, like the people the man met coming from St. Ives, each of these different concepts have associated with them a maximum of 9 numbered narrow concepts.

7 CONCEPT

1. NARROW CONCEPT
2. NARROW CONCEPT
3. NARROW CONCEPT
4. } RESERVED FOR FUTURE USE
5. }
6. }
7. }
8. }
9. OTHER

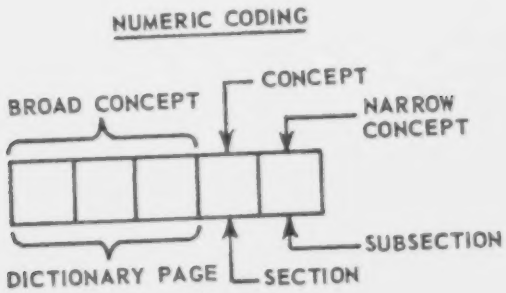
8 CONCEPT

1. NARROW CONCEPT

SLIDE 5

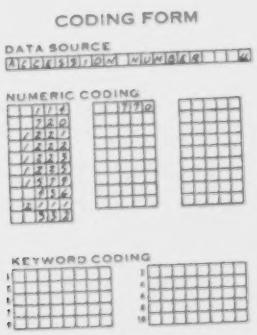
If I may draw my example from my current reading interest, the broad concept or page title could be Fictional Characters, the Concept Animal like, and the narrow concepts – unicorn, dragon, hobbit and the like.

The 5 digit code then, as shown on the next slide,



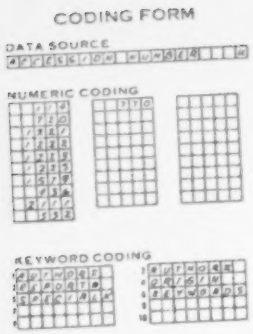
SLIDE 6

gives the page (thousands and hundreds digits), the section (tens digit), and the subsection (units digit) to define the broad concept, the concept, and the narrow concept. It can easily be envisaged that multiple coding from a carefully devised concept dictionary will describe information from any source.



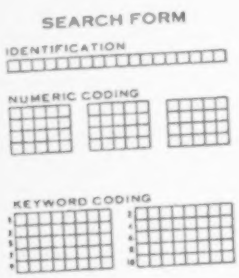
SLIDE 7

But what if a concept which will adequately define a piece of information cannot be found in the concept coding dictionary? What if we want to code "in depth" very narrow concepts, such as a specific chemical compound, an author or a document number? To allow for an open ended dictionary and to accommodate very narrow concepts, an alphameric keyword coding section was included. In our system 10 groups of 8 alphameric symbols can be entered on the coding form.



SLIDE 8

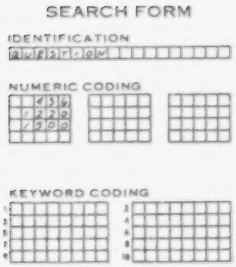
Having considered the input to the system let us examine the means of retrieval.



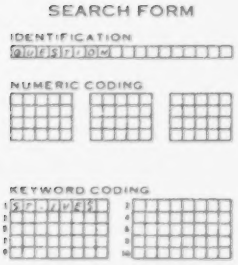
SLIDE 9

The search procedure is just the reverse of coding and depends on a number or letter sequence match by the computer. The next slide shows the Search Coding Form. This form is also divided into three parts. The first is a search identification location to label the question (20 alphameric characters). The second is a numeric coding location to express concepts defined in the dictionary (12-5 digit groups), while the third is an alphameric keyword coding location to define the author, etc., (10-8 character groups). The next slide shows a typical question.

All code numbers and keywords entered on the search coding form are subject to AND logic. For example, a match must be made for all numbers listed. This has not proved to be any great limitation on the system, since our search program has been designed for 10 concurrent searches. Moreover by design, a kind of OR logic has been built into the system so that a broad concept can be searched simply and effectively. The use of 0 in the units column retrieves all documents coded in a concept. In other words, a zero will summon the digits from 1 to 9 inclusive and therefore all narrow concepts in that main concept. Similarly, the use of 0 in the tens and the units column retrieves all documents coded on a given dictionary page. Here the multiple AND becomes OR. There are other interesting ramifications of this system, but before I conclude I would like to describe the control necessary for multiperson use of such a system.



SLIDE 10



SLIDE 11

Many people may be using the same dictionary page for coding and I can avail myself of information my colleagues have coded because all of us use the same dictionary. The insertion of a personal interest dictionary page with coded participants names neatly suffices to separate my file from those of my colleagues. In other words by coding with my number, 323 (page 3 for the user, concept 2 for users names beginning with the letters C and D and narrow concept 3 with my name entered), I can keep my file separate.

Since we live in a throw-away world, we must have some way of throwing away non-current information. To do this, another page of the dictionary is devoted to coding the month and year of origin of the data, and therefore documents published in specific years can be requested. The data base will also be restricted by specifying a particular disc when several of these are filled. At the present rate of use, however, it will take about three years to fill even one disc.

Control over the alphameric keywords is not exercised and by good fortune this facility has not been overtaxed, indicating that the concept dictionary was initially well conceived. A printout of the keywords and the frequency of their occurrence, as well as an excessive use of 9 (reserved for other) in the units column of the numeric coding section points up dictionary weaknesses and has assisted in the consideration of amendments to the dictionary.

The following statistics should speak for themselves.

Number of documents in the file to date:	582
Average number of numbers per document:	10
Average number of keywords per document:	5

On a library card system this means that each document would have to have associated with it some 16 cross-referenced cards.

My thanks go to all those who assisted with the development and implementation of this system: they include about 20 professional and technical officer staff at DRES.

In conclusion, I would like to give you yet another statistic. Average number of times a user has asked for a search to be done in the first ten months of operation: Once! You might say why? Was this experiment a failure? I would have to answer NO for several reasons:

(i) Memory is still fairly acute even after eleven months.

(ii) There have been no extensive program changes at Suffield in the past year requiring literature searches.

(iii) The number of documents in the system is still rather small. This number, however, has recently become sufficient to provide a good data base for questions.

The I.R. dragon is still there but we have begun to find his weaknesses.

EXPERIENCES IN INFORMATION RETRIEVAL USING CHEMICAL ABSTRACTS

by

Dr. R.A. McIvor
DREO

MAGNETIC TAPES

Some years ago, Chemical Abstracts began conversion to an entirely computerized system of production, beginning with indices, and progressing through titles and bibliographic data to complete abstracts. As each step was converted, magnetic tapes were made available to subscribers who wished to operate their own retrieval systems. As an experiment, DRB decided to subscribe to one of these services, which is a subset of Chemical Abstracts covering the interface between Chemistry and Biology. This service is known as Chemical and Biological Activities, or CBAC for short. About 600 journals are covered, in whole or in part, and about 17,000 abstracts per year are included. I might say, in passing, that the entire Chemical Abstracts has about 240,000 abstracts/year. The subscription includes the printed version, which contains a complete print-out of the information on the tapes, ordered by abstract number, as well as a KWIC index of the contents, formula indices, registry number indices, and author indices. The printed version is sold separately, hence there is much indexing that is not needed by the tape user.

DCBRE was given the responsibility of providing information retrieval from CBAC to all DRB personnel wishing it. Chemical Abstracts provide search programs for IBM 360 series or 1401 series computers but no others. In order to use the CDC 3200 computer at Shirley Bay, it was first necessary to write a complete set of search programs. The programs developed are able to search in one pass at least 50 profiles, and probably considerably more, depending on their length. The output can be lists of abstract numbers with title and complete journal reference, or the same with the abstract included, but since the printed abstracts are available, we discourage the use of the latter search mode.

The search programs have the same characteristics as those provided by Chemical Abstracts, so their instructions on profile preparation, and profile forms can be used. I would now like to describe these forms briefly (Slide 1 and 2).

HEADER CARD

QUESTION NUMBER		OUTPUT FORMAT	QUESTION WEIGHT	QUESTION
1		0	25	GATTO: INFORMATION ON AFRICAN LIONS

Fig. 1.

DETAIL CARD

QUESTION NUMBER	PARAMETER NUMBER	NUMBER OF PARAMETERS	MODE	TYPE	LOGIC	TERM	TERM WEIGHT
1	1	2	3	T	O	LION	020

Fig. 2.

The profile consists of two parts – a header card, and detail cards. The header card contains a question number, which serves to identify the profile to the search routine, the output format desired (titles only, or titles and abstract), question weight, which is the minimum weight necessary for retrieval of a record, and a descriptive sentence, which identifies the profile to the human distributing the output.

MODES

(1)	PREFIX	ACTIVity
(2)	SUFFIX	methylAMINE
(3)	WORD OR PHRASE	ACTIVE SITE
(4)	INFIX	2-chloroMETHYLamine

Fig. 3.

TYPES

A	AUTHOR
C	CODEN
T	TITLE OR TEXT
M	MOLECULAR FORMULA
R	REGISTRY NUMBER

Fig. 4.

The detail cards may be as many as desired, one for each term to be searched. The terms can be phrases, single words, or parts of words, and each term has a mode (Slide 3). A group of terms together may make a parameter. The criterion for retrieval is the presence of at least one term from each parameter. In addition, each term has a type (Slide 4) and a term weight. As terms are located, the weights are summed, one weight for each occurrence of the term. The sum is checked to see if it exceeds the question weight and the output is ordered so that within each profile reports with larger weights are listed first. Weights may be positive or negative, and not logic can also be used – that is, a report is not retrieved, no matter what the weight, if a not term is present.

As the information on the tape is in American English, care must be taken to assure that terms are properly spelled, or in case of doubt, all spellings are included. The occurrence of grammatical inflections must be allowed for, either by the use of multiple listing, or by use of the prefix or suffix mode.

The principal difficulty in the use of such a system is anticipating what words will be used by the author, and not by other authors writing on unrelated topics. Let us imagine one wished to retrieve all information about African lions in a collection of papers on general topics, but using this search system. One might use a profile like the one in the next slide.

PARAMETER	TERM	WEIGHT
1	LION*	25
1	FELIS LEO	25
1	FELI*	15
1	*CAT*	15
1	*MAMMAL*	5
1	WILDLIFE	5
1	GAME	5
1	ANIMAL*	3
2	AFRICA*	5
QUESTION WEIGHT		25

Fig. 5.

On submitting this, one might be astonished to receive only one or two retrievals as on the next slide;

EXTENT OF DUPLICATION OF RESEARCH IN THE TRANSVAAL
TH VAN NIETS B.S. DERIEN

CHYMOTRYPSIN ESTERASE ACTIVITY IN GLOSSINA MORSITANS
AHMED IBN KALBAT

Fig. 6.

and this despite his knowledge that many relevant papers were present. The reason for the first retrieval is evident – cat is found in duplication, and Africa was probably mentioned in the abstract, since the Transvaal is involved. The second one is completely mystifying at first, until on studying the abstract we discover that location was mentioned, and since the paper deals with tsetse flies, Africa is mentioned in the abstract. Apparently one parameter only should have been used, as the requisite papers did not mention both African and lion in the title or abstract.

Including both in the same parameter, and changing cat to a prefix instead of an infix, so as to get both cat and cats, but not, for example wildcat or bobcat, we now get a much larger retrieval, a portion of which is seen on the next slide.

BORN FREE
J. ADAMSON

GREAT RELIGIONS OF THE WORLD
F. GOTTDENKER

REDOX PROCESSES AT A METAL CATHODE
S.C.H. WINDLER

A BAYESIAN LOOK AT THE STANLEY CUP
B.E. RODDEN

Fig. 7.

The first indicates that even a fictitious profile can give a relevant item. In the next, of course, Catholicism and the Roman Catholic Church were mentioned more than once. The third mentioned cathodes and cations while the last is about our Great Canadian Game. The next slide gives another example.

CATASTROPHIC EFFECTS OF INJUDICIOUS SELECTION OF TRUNCATION TERMS
R.A. McIVOR

Fig. 8.

Eventually our hypothetical profile writer decided to list both CAT and CATS as complete words, and to delete game as redundant.

The above illustrates the typical problems that arise in the use of CBAC. Usually the best plan is to err initially on the side of completeness, and by examining the undesired retrieval find which terms are causing it and modify them. It is easy to overlook unexpected meanings or word completions. A few examples are given in the next slides.

ELIMINATED *ECTION
REDUCED OUTPUT 20-FOLD

Fig. 9.

In the last of these, the user probably thought he would kill several birds with one stone – detection, protection, collection, and perhaps infection, but overlooked inflection, selection, defection, perfection, convection, direction, correction, section, and dissection, all of which might commonly occur on the tapes.

Currently we have 36 profiles in the system with about 550 terms, ranging from one profile with only one, to two with about 60 terms each depending on the preference of the user for very general words or very specific ones. The users consist of 15 chemists and 8 biologists. Some users have more than one profile.

The reaction of most users has been favorable but one or two have been discouraged by irrelevant output, and others by incomplete coverage of their interests. Two users have asked to be discontinued. People that are most pleased (7 chemists, 3 biologists) have interests that overlap considering the type of material to be found on the tape.

Six chemists and 3 biologists find the service virtually useless, though approving of the principle. Their interests are served to such a slight extent by the topics that they must rely on conventional literature searching for current awareness, and the additional effort required to cover the small fraction of their interests included on the tapes is minimal. Two chemists and 2 biologists have not commented, but I expect they will be more inclined to the second opinion than the first.

Percentage relevance of retrieved article in our experiment still varies widely. Figs. 10, 11, 12, 13, show examples of improving relevance with restructuring of search. The Chemical Society has made a detailed study of 140 CBAC profiles, and found that after optimization, an average relevance of 40 per cent was obtained. This corresponds closely to the best results obtained at DCBRE.

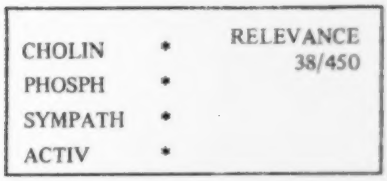


Fig. 10.

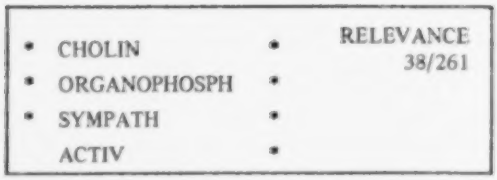


Fig. 11.

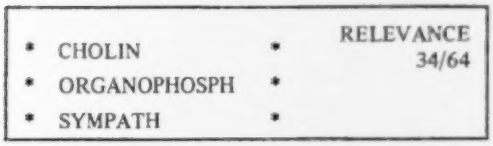


Fig. 12.

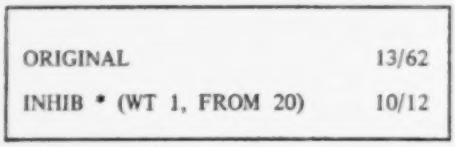


Fig. 13.

In summary, it appears that the computer can be a useful adjunct in informing research workers of current research relative to their problems, but that to have a wide appeal to DRB professionals, a much wider coverage of current journals would be required.

SOCRATES – AN INFORMATION SYSTEM FOR THE CANADIAN DEFENCE COMMUNITY

by

Dr. G.X. Amey

The Defence Scientific Information Service (DSIS) has provided an excellent documentation system to the Canadian defence community since it was inaugurated in 1947. It was the first organization to supply information to its users according to their specialized needs, routinely and on a national scale. DSIS has been a leader in developing the concept of the 'information scientist' – i.e., a seasoned scientist who is devoted full-time to developing expertise in acquiring information in his own field of scientific competence.

Largely because of the excellence of the people-based organization, DSIS did not feel the need for introducing automation until the advent of this current generation of computers. Now, for the first time systems are available that allow many scientists to communicate simultaneously with the data bank of a computer, each receiving rapid response to his queries, as if he were the only user of the system. It is also possible to store huge amounts of information economically and yet have direct and rapid access to all items. Since the flow of scientific information increases steadily with each year, and conventional systems are not adapted for easy weeding of obsolete material, it is more difficult each year to find particular documents in the accumulated material. Traditionally the only solution is to add more staff to cope with the uncontrollably increasing load. The opportune arrival of the third generation of computers seemed to offer a solution to handling the flood of information without indefinitely increasing staff or decreasing the quality of service.

The introduction of a first computer system is always accompanied by an abrupt increase in capital expenditure. This encourages the analysis of traditional and new systems in terms of cost effectiveness, usually for the first time in a library-type organization.

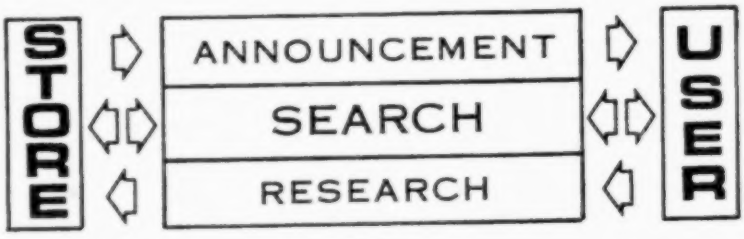


Fig. 1.

Fig. 1 shows the channels of communication between the source of information and the user. The answer to a problem may be found either by undertaking research or by searching the store of available knowledge. Literature searches are expensive. It is usually better to obtain information by dialogue – Socratic dialogue – between an expert and the man with the problem. The expert is kept continuously up-to-date by means of the third channel shown in the figure, i.e., by announcement of new information in his field as soon as it has been entered in the system. This is known as Selective Dissemination of Information (SDI), and in a sophisticated system such as that of DSIS, all new documents are announced selectively and rapidly to all scientists in the system.

We shall now consider the economic basis on which the Defence Research Board supplies information to the Government and the defence community. The principal outputs are consultant services and scientific papers. If we take the total output to be the equivalent of 1,000 papers per annum, costing about $25,000 per paper, this accounts for about half the DRB budget, the rest being accounted for by specific development projects in which either hardware or design data is the chief outcome.

Inputs of data to the system comprise 30,000 reports annually in the form of hard copy documents or microfiche, and perhaps 500,000 papers in the open literature handled by the establishment librarians. Access to the government reports is rigorously controlled by DSIS and the total cost of the operation is less than $1,000,000 per annum, to catalog, announce and control the distribution of documents resulting from research costing the originators about $1,000,000,000. In this estimate an allowance was made also for handling a portion of the accumulated 400,000 documents in DSIS holdings, which store the results of research costing about ten billion dollars. A comparable organization to DRB, the Battelle Research Institute, has formalized the provision of information on a commercial basis through the medium of its Information Analysis Centers. Recently a spokesman for Battelle stated that the income from these services was greater than that from research proper.

On the face of it, money spent on retrieval from world sources of information is much more effectively used than that spent on research in the laboratory. This is certainly true when insufficient money is available to ensure significant results in a field already being vigorously exploited elsewhere. Internal research is imperative where the area involved is very sensitive, where local requirements are unique and the local team is uniquely capable in this field, and where an adequate level of competence must be developed so that the results of research by other workers can be fully exploited. In the area of controlled literature, there is another reason illustrated in Fig. 2. Research may be carried out to produce papers that serve as a bargaining counter in dealing with other research workers. One cannot draw indefinitely from world stores of information without making a fair return. Two extreme situations are shown in the figure, which depend on whether the participants in the exchange are net suppliers (sources) or net absorbers (sinks) of information. The net supplier is defined as an organization that transmits more papers to an organization than it receives in return, and is typified by the United States of America. Canada is a net absorber of information, as it generates only a few per cent of its own information needs through research, both within and outside the defence field.

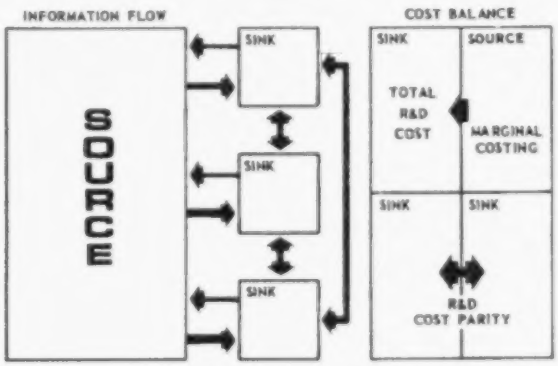


Fig. 2. Economics of Information Exchange.

It appears that a net supplier such as the USA attempts to generate practically all its requirements for scientific and technical information (STI) through research, and thus in determining the cost of information supplied to information sinks, can use marginal cost accounting techniques. Thus in formulating a conceptual model for information exchange between sources and sinks, we may equate the marginal cost of information directed out from the net supplier to the total cost of R&D represented by papers from a net absorber. In the case of two net absorbers we would expect strict parity in costs. The optimum effectiveness of the system will be achieved when sufficient funds are transferred from laboratory research to ensure that, without reducing internal R&D to the point that outside sources are substantially reduced, answers to problems are obtained by searching available stores of STI rather than by conducting laboratory research. This requires that in any individual case, the cost of a search be no more than the product of the cost of research times the probability of finding an answer to the question. Thus, if the probability of finding the answer in the existing data bank is 1 per cent, then a search should not cost more than $250 if it is to be as cost effective as research costing $25,000. Since searching a data bank is usually much quicker than conducting research, it is to be preferred in those cases where time is of the essence.

An economic analysis may also be made of the cost effectiveness of individual scientists as they progress through their normal life-cycle. As the scientist matures, his creativity often falls even as his experience and critical faculties improve. Some gravitate to managerial work or teaching. Others will find that consulting or information analysis are more rewarding. Information analysis and consulting are carried out informally by all scientists of DRB, but it might be well to formalize such services, perhaps as Battelle has done. Output of research papers relative to budget gives a measure of the productivity of a generator of STI. The cost effectiveness of information analysts can be related to their skill in finding answers to problems, within existing scientific literature.

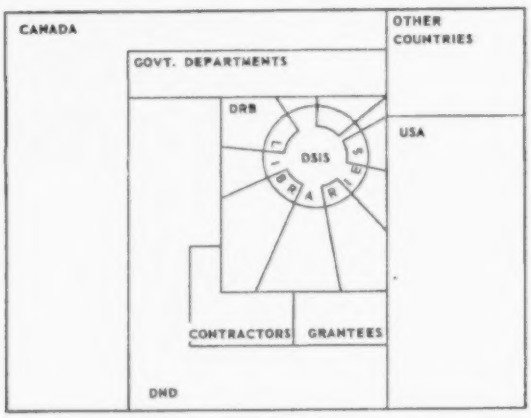


Fig. 3.

In Fig. 3 we see some indication of the complexity of the problem of obtaining information from available sources. The entire field represents the world. STI from the open literature may be channeled to users through the National Science Library, other government departments, or DSIS. The prime responsibility of DSIS is to control the movement of defence-oriented reports from the sources to the users. The range of users is of course coextensive with that of sources, as any scientist fulfills both roles. In terms of quantity of information, our main users are the Canadian forces and their contractors, apart from the scientists of DRB.

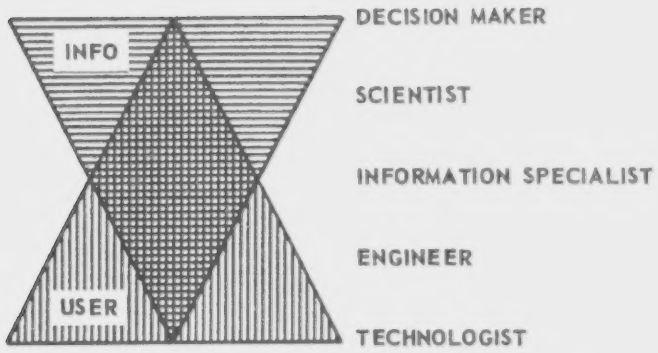


Fig. 4. User Hierarchy.

In Fig. 4 is shown an attempt to categorize users according to their level within the scientific hierarchy. There are few decision-makers, as shown at the top of one pyramid, but their needs touch on every area of knowledge. The quantity of information is shown by the inverted pyramid. Technicians, on the other hand, are many but their STI needs relatively simple.

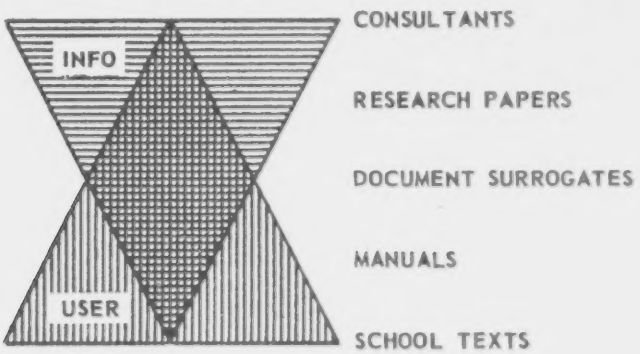


Fig. 5. Communication Media.

Fig. 5 is the complement of Fig. 4 and indicates the variation in media appropriate for communicating with various classes of user. In the past, DSIS has concentrated on the class of working scientists for whom research papers are grist to the mill. The principal medium of communication for the information scientists consists of document surrogates, i.e., indexing data and abstracts that can serve as substitutes for documents when conducting searches of the literature.

The newest aids in information retrieval are surrogates in machine-readable form, and documents in microform. By the use of computerized data banks developed from bibliographic data on magnetic tape, searches can be conducted rapidly and from remote locations. The microfiche is much quicker and cheaper to reproduce than hard copy and thus speeds up delivery of a document selected as the result of a search. Formerly reproduction of hard copy documents was carried out only rarely, because of the cost. When a copy was not available at DSIS, long delays could occur while it was being obtained from the holder.

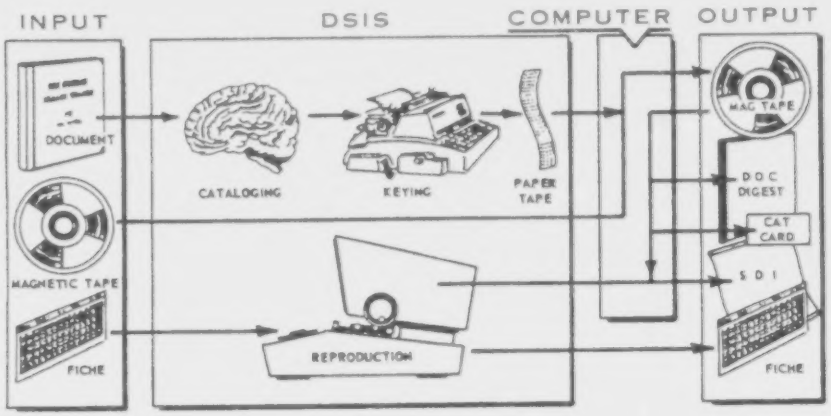


Fig. 6. SOCRATES Document Processing.

Fig. 6 contrasts the traditional system with the computerized system being developed under project SOCRATES – System for Organizing Current Reports to Aid TEchnology and Science. (Ref. 2, 3). In the upper part of the figure, the operations involved in preparing a catalog card and announcement bulletin are shown for the case where surrogate data on magnetic tape is not available. Each library entering the document into the system writes an abstract, selects indexing terms, and keys the data into its own system. Since 1 January, 1967, we have used a flexowriter for this purpose. In addition to the typed outputs so obtained, a byproduct paper tape is available containing machine-readable data for entry to a computerized data bank. In this case, each library receiving the document duplicates the cataloging process. When the surrogate is available on magnetic tape, this complex input stage, which costs more than $10 per document, is replaced by the operation of translating from one code to another. In the case of our Unitran program, this requires six seconds or 20¢ per document.

The DRB information network from the worm's eye view of an individual user is shown in Fig. 7. Heavy lines indicate those cases where cataloging information is available on magnetic tape, rather than only on printed media. At present the formats used for transmitting such data have not been standardized so that the user must treat each service on an individual basis. There are several hundred such in existence. The need for a capping agency within the network to undertake the task of standardization is evident. If the ideal is to be realized, that a scientist will do a thorough search of the literature before undertaking costly research, it would appear that all relevant material whether in the open or the controlled areas, should be accessible on a uniform basis, instantaneously, from any part of the country. At present, each scientist must shift for himself. It has been reported that the habits of scientists under this regime extend from those who read no journals or only one, and seek STI only by verbal or phone contacts with their peers, to one individual who steadily combs through the entire chemical literature, which constitutes about a quarter of a million papers per annum.

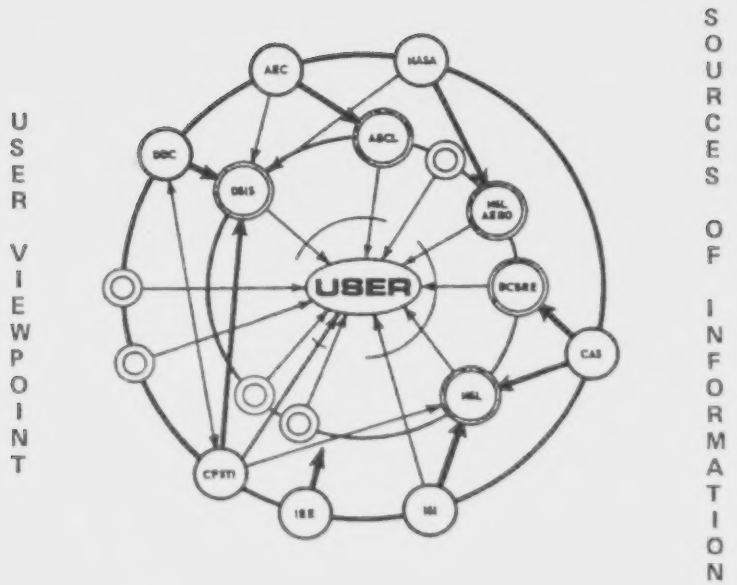


Fig. 7.

In the course of the SOCRATES project, we are looking at the total information problem on the one hand, and on the other hand developing a computerized data bank that will be usable by any system we should come up with. Our philosophy in forming this data bank is that all types of data should be entered on a uniform basis, that is, we should not constrain the system so that it is suitable only for filing data about documents.

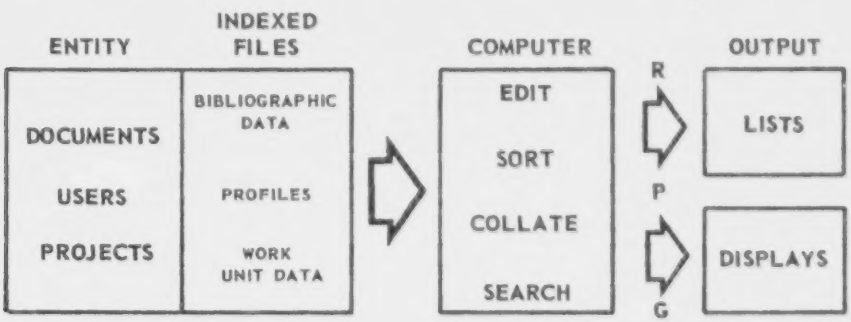


Fig. 8. SOCRATES File Management.

Fig. 8 blocks out the main features of the system. Data that references any entity is entered into linear and structured files associated with directories and dictionaries, which allow of compact storage, easy update and rapid retrieval. These files can be manipulated to provide ordered outputs according to the needs and type of user, either printed or on CRT displays. It is hoped that the variety of outputs, and the command language

used in communicating with the system will be such as to attract users higher and lower in the hierarchy than current DSIS services do. For this purpose, access through natural language (French and English) will be attempted, special program packages can be developed to display work unit data, and initiation to the system will be via computer-aided instruction.

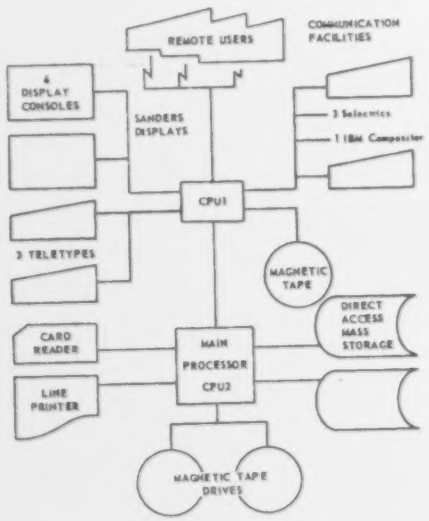


Fig. 9. Hardware Configuration 1971–72.

Fig. 9 shows the proposed computer configuration being built up in accordance with a five-year plan. We have been keying the data in the data-bank for nearly two years and have been producing one weekly announcement bulletin sinch March of this year. All printed outputs of the traditional system will be obtained via the computer by 1 January, 1969 and in addition they will be improved in format, indexed and cumulated as never before. With the addition of the small computer, CPU1, we shall be able to enter data online to magnetic tape next year. Via the teletypes it will be possible to control circulation of documents by a hybrid system of man and machine. Until 1970, most operations involving the computer will be carried out in the batched mode. On receipt of the large computer, CPU2, the various tasks will be brought into online operation. Fig. 10 shows projected changes in staff as the new system is phased in. Phasing out of the old operation will require parallel operation of both systems for a number of years and will delay projected savings. As shown in the figure, although current plans should be implemented by 1972, phasing out of the old system will probably take until 1975. We shall then have a system much closer to the optimum than the present one, from the point of view of clerical operations, and able to handle the open literature if this is so desired. The effectiveness of the total system will however, depend on the use made of the stored data and the extent to which a formal structure appropriate for comprehensive information analysis is developed.

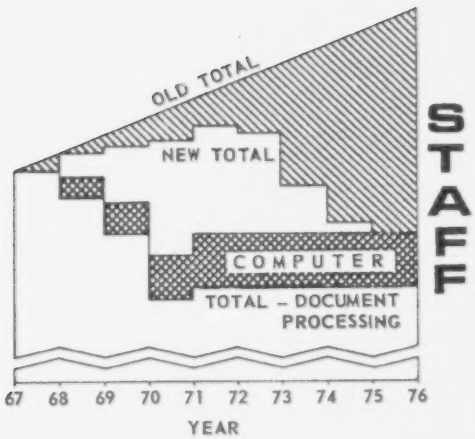


Fig. 10. DSIS Staff.

REFERENCES

1. G.X. Amey: "Channel Hierarchies for Matching Information Sources to User's Needs". Proceedings of the American Society for Information Science, Vol. 5 (1968).

2. G.X. Amey: "The Evolving Defence Scientific Information System". Proc. of the 2nd DRB Librarians' Conference. (1967).

3. G.X. Amey: "Evolutionary Transition to Automation of Information Management". Proc. of the 5th Colloquium on Information Retrieval. (1968).

FILE DESIGN FOR PROJECT SOCRATES

by

J.G. Moxham

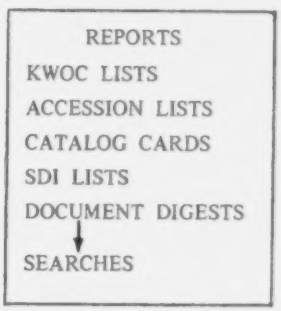
INTRODUCTION

Before delving into the details of File Design for PROJECT SOCRATES, it should be pointed out that time will limit the scope of the paper to a reasonably narrow area of DSIS activity. The discussion will dwell mainly on the present state of the system and its development in the very near future, and will end with a brief look at the problem of putting the system on-line.

Although we are developing our system with the capability of handling many types of files, the remainder of the paper will be concerned mainly with how Bibliographic data files are handled.

A brief introduction to the system is necessary and this will start with a discussion of what output System SOCRATES is, or will soon be, capable of producing.

SLIDE 1



At the moment, the system is only capable of producing reports of a routine nature. Catalog Cards and Document Digests are not new and were both produced by the manual system.

KWOC Lists (Key Word Out of Context) and Accession Lists are the most important new system outputs now in production and certainly require the use of a computer for economic production.

SDI (Selective Dissemination of Information) List programs are in the process of being implemented (SDI lists are being produced experimentally at the moment). This part of the system will allow the scientist to define an interest profile for himself using sets of Boolean relationships of indexing terms. As a result he will receive, regularly, notice of new documents in his own field of interest.

It should be pointed out that all the above reports are produced regularly on new bibliographies entering the system. At the same time, a data bank is being created that will allow SEARCHES to be performed on the whole data bank (not only on new entries). The technique for Searching will initially be roughly the same as for SDI, but the profiles will have to be more narrowly defined. Searches will be done on a batch basis, and turn-around time will be much slower than with a real time system.

The system under discussion is being designed around the CDC 3200 computer at DRTE. Input, at the moment, is by means of paper tape produced on Flexowriters. Within the next year, the input medium will be upgraded to magnetic tape to be produced on a small computer with multiple typewriter input.

CODING OF DOCUMENTS

Information being stored in our data bank is of surprising variety. In order to classify this variety of information, a unique code is used to identify each type of bibliographic information. This consists of a two-character field identifier allowing for up to 99 fields. Within each of these fields a further optional two-character identifier is provided to allow a further breakdown of fields into a possible associated 99 subfields.

SLIDE 2

BROAD CATEGORIES
ACCESSION NUMBER
SECURITY & SENSITIVITY
TITLE
AUTHORS
ASSOCIATED PROJECTS, ETC.
DATES
ABSTRACT
DESCRIPTORS, KEYWORDS, ETC.

Information represented in our bibliographic data bank may be broadly categorized as shown on this slide.

- **Accession Number** is a unique number allocated to each document entering DSIS.
- **Security and Sensitivity** are self-explanatory and are used by the system to control access to documents by unauthorized personnel.
- **Dates** cover such data as publication date and period covered by the document.
- **Title** does not have to be described.
- **Authors** includes all involved in preparation of the document, including Corporate sources (Organizations supporting or originating the report).
- **Associated Projects** catalogs, under various sub-headings, project and task numbers allocated to the report or its topic, by various agencies such as DRB.

— **Abstracts** are originated either by the author or Information Scientists either in DSIS or in other similar agencies supplying information to DSIS. These are usually sufficiently condensed to be printed completely on the catalog card.

— **Descriptors** are undoubtedly the most important bibliographic information stored on documents. These are supplied at various stages by information scientists (intellectually generated descriptors) and, in the case of this system, will include automatically generated key words from titles as well as more traditional subject headings or terms extracted from various Thesauri and broad categorizations such as COSATI codes. Descriptors are expected to be the most common elements used for automated document searches as the system develops.

THE COMPUTER SOFTWARE PROBLEM

It has now been established that certain information is gathered in computer-compatible form for creation of a data bank of document bibliographies.

From a computer software viewpoint, some problems now become apparent mainly owing to the variability in the nature of data being stored. These problems may be summarized as follows:

(a) First, a problem is introduced by the fact that bibliographic information on practically every document requires a different combination of fields.

(b) Secondly, once the required fields have been established, we find that there is an immense variation in the length of data stored in each field.

SLIDE 3

TITLES
04 CAMOUFLAGE PRACTICES
04 POLAR CAP ABSORPTIONS AND ASSOCIATED SOLAR-TERRESTRIAL EVENTS THROUGHOUT THE 19TH SOLAR CYCLES

This slide illustrates the problem by showing two titles (titles use field number 04.) We obviously cannot allocate enough fixed storage in our title field to cope with the largest title expected.

(c) Thirdly, there is a problem introduced by the fact that some fields may sometimes contain a variable number of sub-fields.

SLIDE 4

AUTHORS
11 SELL, R.E.
11 MERRICK, V.K. MORAN, F.J. TINLING, B.E.

This is illustrated by the case of two documents, one originated by a single author, the other a joint effort by three authors. We obviously cannot allocate a sufficiently large number of sub-fields to cope with the largest number of authors expected.

We can judge from the above examples that our files cannot be of fixed size with a predetermined area allocated to each field and with blank space wasted for unused fields. This approach would result in immense records containing mostly nothing.

Our files must, then, be constructed to store variable-length data compactly. Another constraint that must be mentioned is that records must be constructed in such a way as to allow access to any field or sub-field with a minimum amount of computer search time.

The most obvious way of attacking the problem is to generate a continuous string of information for each bibliography with some sort of unique identifiers at the ends of each unique information field. This method is, in fact, being used at the moment. The technique is an inheritance from a previous system design. The obvious flaw in the design is that searching for information is very time-consuming, as all information up to the required field must be scanned on a character-by-character basis to find what is required. Once the information has been found, the search continues to establish where the required field ends.

The file design to be described is an effort to eliminate these problems. It is to be used initially for a magnetic tape system using the DRTE CDC 3200 computer, but will eventually be converted to a random access storage medium. Records are organized to allow a binary search technique to be used within each record (one bibliography).

THE FILE DESIGN

The structure of the file is such that each record can be broken into three basic components. See Figure 1 (substituted for slides 5 to 8).

(a) **A Fixed Data Area**, which contains a small amount of information common in length to all records in the file. This area also contains some control information.

(b) **A Directory Area**, which relates field numbers to where the field is stored in the Text Entry area (see c).

(c) **A Text and Dictionary Entry Area**, which is stored more or less as a continuous string of information, each piece of which can be located with the aid of the Directory Area.

FIELD NAME	Fixed or Variable	Char or Binary	Usual No. of Chars.	Remarks
FIXED DATA [a][b][c][d][e]				
(a) Length Word	F	B	4	Word count of whole record (a) included
(b) Accession Number	F	B	4	dash (–) excluded
(c) Status Word	F	B	4	bit oriented
(d) Security	F	C	1	
(e) Sensitivity	V	C	1 – 4	characters left justified
DIRECTORY [f][g_1][g_2][g_3][g_4]{[g_n][h_1]…[h_n]}				
(f) Error Area Pointer	F	B	4	Words relative to (a) 0 if no errors.
(g) Field Nos. & Pointers	F	B	4	Pointers to Text – one (g) entry per field number.
(h) Error F No. & Pointers	F	B	4	(see g) One (f) entry per error field
TEXT & DICTIONARIES [i_1][j_1]{[i_2][j_1]}{[i_n][j_n]}				
(i) No. of Sub-fields	F	B	4	Missing if no sub-fields, otherwise one (i) entry per field
(j) 1. Text	V	C	1 – 1000	Sub-fields transmitted by
2. Dictionary	F	B	4	

Fig. 1. Record Layout of Bibliographic File.

These areas will now be described in greater detail.

The FIXED DATA area consists of five items:

(a) a length word, which defines the total length of the record,

(b) the accession number, which is a unique number for each document,

(c) the status word, which controls production of reports and retention of information on files (this will be enlarged upon later),

(d) & (e) Security and Sensitivity codes, which are the key to accessing information on the remainder of the record by the users. This will be especially useful when on-line searches are incorporated in the system.

At a later date when the files are placed on random access storage, another field may be added as a pointer to the start of the remainder of the information in the record. At this point the Fixed Data entries will be separated from the remainder of the record.

The DIRECTORY is divided roughly into two parts:

(1) An ordered set of field numbers for Text & Dictionary Entries that passed into the file through a computer edit without any faults being detected. These are designated g_1 to g_n on the slide.

(2) An ordered set of fields (designated h_1 to h_n) that have been accepted into the file by the edit but are known to contain errors. These fields may be modified by inputting changes to the system.

Both of these strings can be searched easily. Each entry in both fields is divided into a field number and its relative address in the Text Area. The entry designated f points to the start of the error area. Zero indicates that no errors are present on the record.

Each field in the TEXT and DICTIONARY area is headed by a count of the number of sub-fields in the field (if any can be present). Otherwise this is missing. The sub-fields are variable length except in the case of Dictionary Addresses. The program is expected to be aware of which fields may contain sub-fields and which sub-fields are dictionary entries.

Most files envisaged by DSIS at the moment will fit into the structure just described. Files planned for future inclusion are Patron, Project and Loan files, but this need not set the limit.

In manipulating the files described above, it is desirable that they do not become too bulky. A double file system is planned – one for recent additions covering roughly the latest two months of data from which reports are still being produced – a second, and much larger, file is planned for all other bibliographies. A periodic computer run will be carried out to move data from the small to the large file. This operation makes use of the Status Word mentioned above, which is changed by each report produced and by changes in record content such as elimination of all error fields.

KEY WORD SELECTION

One file, not yet described, is that associated with KWOC List production. Keywords for the KWOC Lists are selected by running each word in the title against a stop list – a list of words not desired as Keywords.

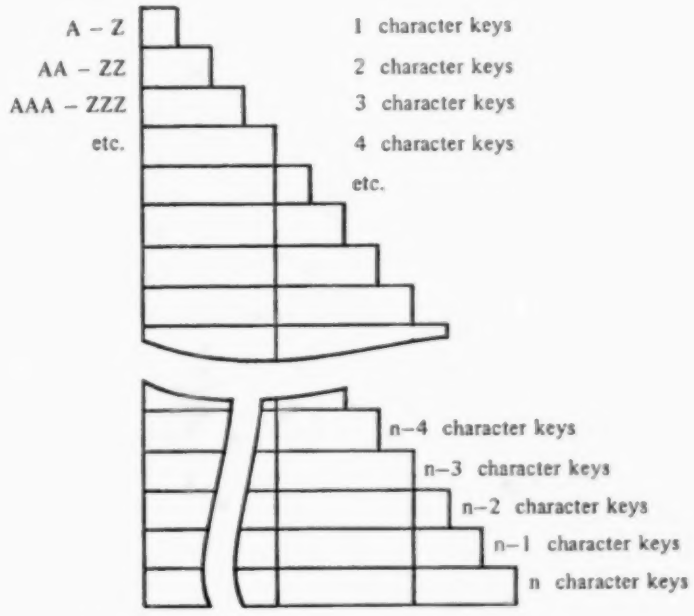
SLIDE 9

STOP WORDS	ANTI-STOP WORDS
MATTER/	
MEANS	
MEASUR/	AIR
MECHANISM/	CAS
MEDIA	ION
MEETING/	
MEMBER/	
METHOD/	
MIDCOURSE/	
MIDDLE	
MILITARY	
MILLIMETER/	
MINIMUM	
MINUT/	

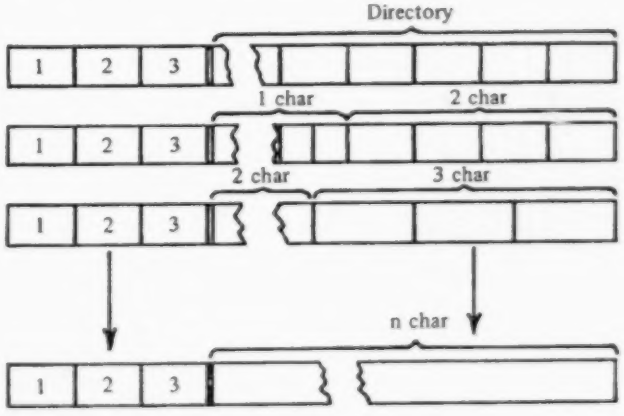
This slide shows (on the left) an excerpt from the present Stop List. Slashes at the end of some of the stop words indicate that these are root words and will cause rejection of any Title word beginning with this root. The present method used in scanning for Stop Words is essentially one of matching each title word, character-by-character, with the alphabetically ordered Stop List – not the most efficient method for a word-oriented computer.

In an effort to speed up the operation and to reduce the number of Stop Words, all words of three letters or less were considered as Stop Words. However, exceptions to this rule soon became apparent creating the need for an anti-Stop List. The only entries yet appearing in this list are shown at the right of the slide.

In an effort to speed up the Stop List scanning procedure, a restructuring of the list is to be implemented. See Figure 2 (substituted for Slide 10).



(a) ORGANIZATION IN COMPUTER



(1) Length of Word
(2) Sequencing Key
(3) Dummy Status Word

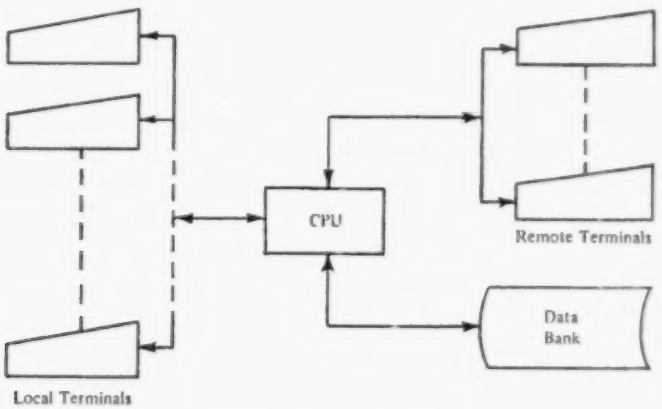
(b) MAGNETIC TAPE FILE ORGANIZATION

Fig. 2. Stop and Anti-Stop Lists

With the set-up shown on this slide, Stop Words will be grouped by word size (natural language) and ordered by the integer value of the computer words thus generated. Integer Binary Search techniques will then be possible within word size groups if a directory is provided for the start of each of these groups. Of course, more than one group must be searched to cope with root Stop Words. Anti-Stop Words will also be included in this file.

FUTURE SYSTEMS

The system is eventually intended to evolve into a real-time system. Some aspects of this future system are briefly described below.

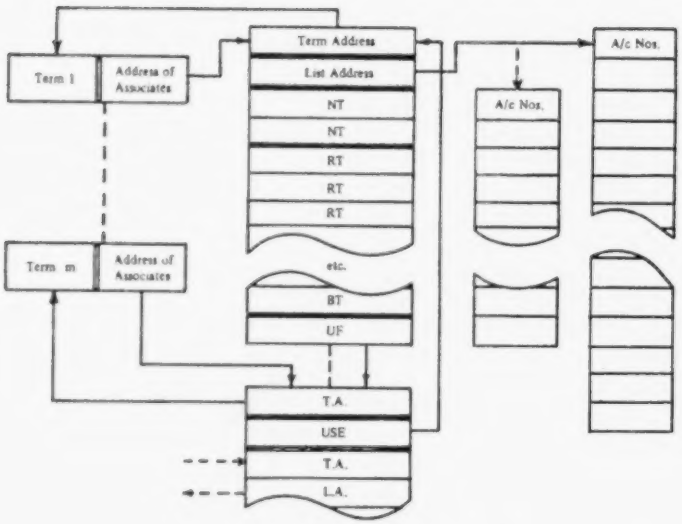


SYSTEM CONFIGURATION

SLIDE 11

This slide depicts a computer with a data bank onto which is hooked various terminal devices. The majority of terminals are expected to be local because of high line costs. The system is intended to be conversational to the point where a relatively inexperienced user or Information Scientist may use the system with a 'helping hand' from the computer. The data bank consists of a large bibliographic data base with various index files.

We will discuss one possible file organization that could be used with this conversational system.



ON-LINE SEARCHING

SLIDE 12

This set-up uses a mass storage system. The left hand side of the slide is intended to depict a thesaurus stored in computer form. Assume that its terms are organized in the same form as the Stop List previously described to allow for binary searching. The right hand side represents ordered lists of documents using the various descriptors appearing in the thesaurus.

It will be noted that each thesaurus term has tagged onto it a pointer to a variable length list of Associated Entries as in a natural language thesaurus. The list includes all the normal thesaurus relationships such as USE, UF, RT, NT and BT references. These are, however, all stored in the form of further pointers to similar lists.

The first entry (the Term Address) in each list of associated terms may not at first appear too sensible – it points back to the term that pointed to it. The reason will become more obvious later.

To illustrate the use of this pointer system, let us trace what happens when a user wishes to get to the list of documents using term 'm' as a descriptor.

(1) We find that there is a preferred term (USE) in the Associated Term list for term m, which points to another term list.

(2) The reason for Term Address now becomes apparent. It points to the natural language version of the preferred term, and the computer may type it out for the user if required.

(3) The second entry (the List Address) points to the list of associated document accession numbers, which is again ordered numerically. These lists or inverted files serve two purposes:

(a) they list all accession numbers using this term,

(b) they can be used to find the complete bibliography, which is also stored in our bulk file. A simple algorithm can be used to compute the address of the bibliographies associated with accession numbers in the inverted files.

It can be seen that these files may have additional entries for keeping statistical information on the usage of terms, etc.

We now see how the inverted files can be used in Boolean searches. Assume we are trying to find documents that have as descriptors both MYXOMATOSIS and RABBITS. Assume we have been led to the two lists shown on the slide. It is now a matter of comparing both lists and picking out accession numbers to guide us to the bibliographies of all documents using both terms.

This is done with the aid of our fixed data field as previously described (see Figure 1). A similar structure can be used for other inverted files such as Keywords and the various Author Files. These are, however, dynamic files and present an updating problem outside the scope of this paper. The Thesaurus file just described is expected to be revised, roughly, once per year.

In closing, some comments on hardware requirements for our on-line system are in order. Since the operations just described will be in process simultaneously for many users, each desiring responses in the order of 5 seconds, it would be ideal to store inverted files on very fast access mass storage devices (this implies bulk core storage or head per track drum or disc storage). Bibliographies may be stored on slower bulk stores with access times in the 100-200 mS range. Initial storage estimates for 5 years of data place inverted file requirements in the 10 million character range and bibliographic storage at over 100 million characters.

COMMUNICATION FACILITIES FOR AN ON-LINE INFORMATION RETRIEVAL SYSTEM

by

R.W. Grant

INTRODUCTION

A number of communication facilities suited to the requirements of an on-line information retrieval system are available in Canada.

In order to design a suitable network the prime considerations are as follows:

(a) The types of terminals required to handle the volume of traffic and message lengths envisaged.

(b) The number and speed of the communication lines required to handle the traffic.

(c) Communication line layout.

(d) Costs and associated user benefits.

SOCRATES

In this report we will consider the system being developed as part of the SOCRATES project.

This is currently a batch processing system which handles both intellectually and automatically indexed documents. It will soon be developed into an on-line information retrieval system.

Communication networks tend to be extremely costly. The prime considerations for SOCRATES are:

(a) Cost.

(b) The network should be a general purpose one to allow the possibility of inter-establishment communications for administrative messages.

At the moment there is no formal data communication network for SOCRATES' users.

IMPLEMENTATION

A suitable network will be implemented in several broad phases so that the users will be gradually introduced to the facilities of an on-line system.

Statistical data covering such aspects as:

(a) equipment and communication line usage;

(b) retrieval and indexing efficiency;

(c) users reaction to – conversational languages

– terminal facilities

will be gathered and evaluated at each phase of implementation. The results of the analysis will probably influence the development of the latter phases.

The broad phases are shown in Figure 1.

PHASE 1

- BATCHED, OFF-LINE SEARCH
- MANUAL LOAN CONTROL

PHASE 2

- ON-LINE SEARCH
  - □ BATCHED MODE
  - □ CONVERSATIONAL MODE
- ON-LINE LOAN CONTROL

PHASE 3

- CONVERSATIONAL MODE SEARCH
- EXPANDED TERMINAL FACILITIES
- LINKED WITH OTHER INFORMATION RETRIEVAL SYSTEMS

Fig. 1. SOCRATES

IMPLEMENTATION PHASES

Phase 1 — Searches will be performed in an off-line mode with an anticipated response of 12 to 24 hours.

The user search requests will be prepared by an information specialist who is familiar with indexing and the thesaurus used by the system. This will enable accurate and concise search requests to be formed.

The search requests and other messages will be transmitted to a small computer located at DSIS and recorded on magnetic tape. At the end of the day the recorded data will be processed on the main computer and the search results will then be transmitted back to the appropriate establishments.

Phase 2 — In this phase the communication network will be connected directly to the main processing computer at DSIS.

Two modes of searching will be available:

Batched On-line Mode — The same search format as used in Phase 1 will be used, however, the response time will be of the order of minutes. The actual response time experienced will depend on the number and size of searches in a batch and the current loading on the computer.

Conversational Mode — A sample user population will be permitted to perform searches in a man/machine interactive mode. The response time will be of the order of 3 to 10 seconds.

Considerable effort will be made to evaluate the effectiveness of the conversational language provided.

Phase 3 — The conversational mode will be available to all users.

The user terminal facilities will be expanded to include display screens, high-speed printers, magnetic-tape devices etc.

A near English (bilingual) conversational language will be developed.

Communication with other compatible information retrieval systems is foreseen. This could provide users with access to a very extensive range of information.

Later in this phase we may see the introduction of speech input and response.

COMMUNICATION TRAFFIC

The SOCRATES communication traffic was estimated by using data based on the present system and the results of an evaluation of the NASA/RECON prototype information retrieval system.

In order to minimize the transmission costs it was decided that search requests should be batched.

The estimated total daily transmission load between the establishments and DSIS is;

BATCHED SEARCH MODE	40,000 Characters
LOAN CONTROL (BATCHED)	16,000 Characters
CONVERSATIONAL MODE	

The conversational mode loading is not related to the number of characters transmitted and is a function only of the time the circuit is connected (the duration of the conversation). The time required depends on such factors as;

(a) Language design.

(b) Computer and user response times.

(c) Number of errors made.

(d) The success of the search.

However, an allowance of 15 minutes per conversation was made for an average of 5 searches per day.

TERMINAL CHARACTERISTICS

The type of user terminal to meet the above requirements must be;

(a) cheap,

(b) general purpose,

(c) provide a hard copy,

(d) prepare search requests or messages in an off-line mode,

(e) transmit batched searches at maximum speed,

(f) capable of transmitting directly via the keyboard.

It should be noted that the various departments of DSIS will not impose a load on the general network, as their terminals will be connected directly to the central computer. The bulk of the traffic will in fact be internal to DSIS.

COMMUNICATION CHANNEL SPEEDS

The range of communication channel speeds includes:

(a) LOW SPEED – 110 bits/sec. (TWX or DATA TELEX)

(b) VOICE GRADE – 1200 bits/sec.
2000 bits/sec.

(c) PRIVATE LINES – 2400 bits/sec.
4800

Higher speeds are envisaged in the future.

SOCRATES INITIAL COMMUNICATIONS NETWORK

The low-speed channel and the teletypewriter terminal were selected as being most suitable for the initial requirements.

A schematic of the proposed network is shown in Figure 2.

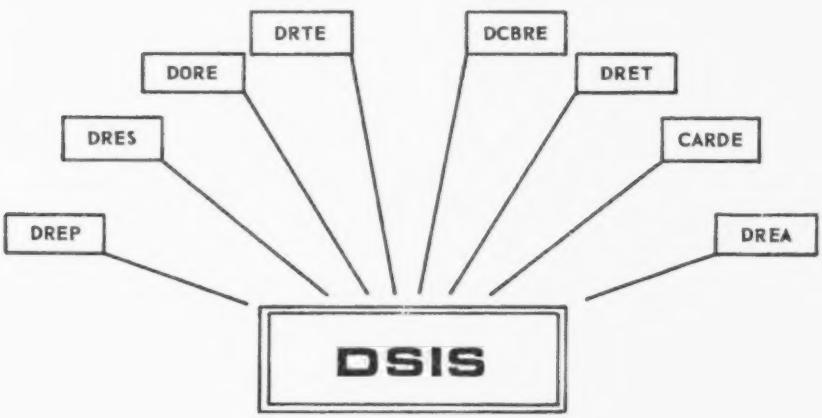


Fig. 2. Basic Communication Network

Establishments are linked directly to each other; however, these lines were omitted from the figure for the sake of clarity.

Each establishment will have either a type 33 or 35 automatic send and receive teletypewriter.

This equipment fulfills all the requirements and can prepare the search requests on paper tape and then transmit them using the paper tape reader at some later time.

These units provide only upper case character typing. Upper and lower case will be available with the model 37 when this becomes available.

Users outside the Defence Research Board will be able to communicate via the general purpose network, subject of course to normal access control in terms of security. (TWX or TELEX)

OPERATING PRINCIPLE

A schematic of the system is shown in Figure 3.

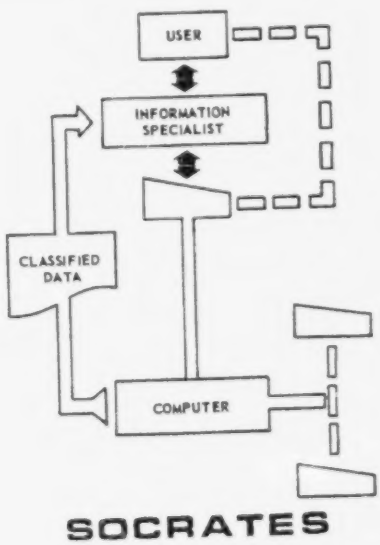


Fig. 3. Phase 2.

The figure shows a single establishment connected to the central computer facility located at DSIS.

The information specialist will assist users in specifying searches or other requests.

Completed requests will then be typed on the teletypewriter and simultaneously punched into paper tape. The messages will be proofread and, where necessary, corrected.

The paper tape containing the messages will be placed in the reading device and transmitted to the computer.

Some users will be permitted to use the equipment for conversational mode searches for experimental purposes. In these cases the user will be directly on-line to the central computer.

On receiving a message, the central computer will perform the necessary functions and store the results of each search. When a batch of messages has been processed, the computer will dial up the appropriate establishment and transmit the results.

Classified data will not be transmitted over the communication network but will be forwarded to each establishment using the conventional channels. Voluminous material will also be handled in this manner to avoid excessive costs.

Message Format

Figure 4 shows a sample search request and the response received from the computer.

```
BATCH
USER            DRTE/RA
                DATA PROCESSING TERMINALS
DATE            AFTER 1967
USER            DRTE/AJL
KEYWORD         CONSOLES
                JOSS
                RECON
USER            DRTE/SN
ORDER           HC,67-07332,68-00231,MF,AD-650 573
                      Etc.
USER            DRTE/RA
SEARCH          1725/1  7  HITS
                DATA PROCESSING TERMINALS,AFTER 1967
CLASSIFIED      68-02334,68-04578,AD-675 415
68-03345        A LOW-COST OUTPUT TERMINAL FOR
  TIME-SHARED COMPUTERS. BY ROSENBERG R.C.
  SOURCE 239110 REPORT NO. MAC-TR-38   Etc.
```

Fig. 4. Search and Response.

The message is formatted with control or key information to the left of the variable search terms.

The control information (BATCH, USER, DATE etc.) is used by the computer to identify the related data appearing to the right.

Taking each line in turn we can understand the method of forming search requests.

Line 1 – The control word BATCH indicates to the computer that the batched mode of operation is desired. The absence of this word would imply conversational mode.

Line 2 – USER specifies that the user's identification code appears on the same line. This example shows the identification as consisting of two parts;

(1) the establishment code (DRTE)

(2) the user initials separated by a slash (1). (RA)

The identification code will be used to determine whether the user code is legal and also the security and sensitivity level of that user.

Line 3 – The absence of control data on this line indicates the specified term (DATA PROCESSING TERMINALS) is a precise thesaurus term.

Line 4 – DATE indicates that the user is interested only in documents printed after the year 1967.

Line 5 – This indicates both the end of the previous search and the start of the next.

Line 6 – KEYWORD specifies that the search is to be performed using keywords. CONSOLES is the first keyword.

Line 7 and 8 indicate further keywords JOSS and RECON.

By entering the keywords on separate lines the user has automatically implied a search logic of CONSOLES OR JOSS OR RECON. The keywords could have been entered in this manner on a single line. Other logical connectives such as AND and NOT are permitted.

Line 9 – Again, this identifies the next user.

Line 10 – ORDER specifies that the user wishes to order the documents specified. HC indicates a hard copy is required. MF indicates a microfiche copy.

The Response

The result of the first search is shown. The user is identified by his unique code.

SEARCH identifies the search performed. The number 1725 is a unique number assigned by the computer. The next number (separated by the slash) indicates that this is the first set of results for this search.

The user can use the search number to modify the terms of the search without having to re-enter the original search terms. In the case where a modified search causes no documents to be retrieved, the user can restart from a previous search point by specifying an appropriate set number. The number of documents retrieved is indicated as 7 HITS in this example.

The terms of the search are printed on the next line.

CLASSIFIED indicates that some classified documents were retrieved as a result of the search. Since classified data will never be transmitted, only the unique document identification (which is unclassified) is given. The user will be able to check the establishment library catalog file for further information on these documents before deciding to order them. Details of any unclassified documents are then printed. There we see the document number, title, author, corporate source code and the report number printed for each document.

The question of data security may be raised at this point. If an illegal user enters another user identity, will he gain access to classified data? As we have seen, classified data will not be transmitted over the network. If the illegal user orders a classified document he must specify a valid user code and in this case the document

would be forwarded to the user specified on the request, not the illegal user. The document is therefore subject to the normal security controls. In the batch mode of operation the search requests will also be subject to the scrutiny of the information specialist.

Conversational Mode Searches – The conversational mode of searching will provide users with continuous interaction with the computer.

The conversational language will enable both experienced and inexperienced users to get maximum benefit from the system. In the case of an experienced user, the computer response will be terse and the user will command the computer, whereas the inexperienced user will basically respond to a number of choices indicated by the computer dialog.

A user will have the ability to search through the thesaurus and select the terms to define the search requirements. As the terms are displayed, the number of documents indexed by that term will also be shown. The search may be broadened or narrowed by selecting appropriate search terms. These may be linked by boolean logic or weighted in any combination to specify an accurate search request.

ANNUAL COMMUNICATION COSTS

The annual communication costs for phase 2 of the system are shown in Figure 5.

COMMUNICATION EQUIPMENT RENTAL (Headquarters and Establishments)	13,140
LINE CHARGES	
BATCHED SEARCH	4,100
CONVERSATIONAL MODE	1,900
LOAN CONTROL (Batched)	1,700
TOTAL	$20,840

Additional cost for conversational mode loan control — $6,600

Fig. 5. Annual Communication Costs.

Communication Equipment Rental: $13,140.

This cost includes the rental for a terminal in each establishment and DSIS as well as the equipment necessary to interface the communication lines to the central computer communications controller.

It does not include the cost of the central computer, the computer message processing cost or the terminal facilities provided for divisions of DSIS.

Line Charges: $7,700

This charge is made up of three entries:

(a) Batched search: 4,100.

(b) Conversational mode: 1,900.

(c) Loan control (Batched): 1,700.

The batched mode of operation minimizes communication costs, as the data transmission speed is limited only by the maximum speed of the equipment. (10 characters per sec.)

The conversational mode charge is much higher, as an average transfer rate using direct keyboard entry is of the order of 2 characters per second when keying data. This is indicated by the increase in costs ($6,600) incurred when loan control is performed in conversational mode.

The line charges do not include the transmission costs for general administrative messages.

The possibility of using the Federal Government or Military networks is also being considered.

USER BENEFITS

The user benefits provided by an on-line information retrieval system will be as follows:

. ability to search large centralized data banks as if they were located on the premises;

. a number of search strategies will be available;

. the facility of displaying document details will enable users to order relevant documents immediately. An immediate response will indicate the action taken. (e.g. where a copy is available etc.);

. the data stored in the bank will be much more up to date than in a batch system, as DSIS operations staff will be employing on-line data entry and editing techniques;

. editing facilities will enable the user to edit the results of a search (i.e. reject irrelevant documents);

. output controls will allow users to direct the search results to selected devices such as.

(a) cathode ray tube displays,

(b) page or high-speed printers,

(c) magnetic tape recording devices,

(d) local computer storage etc.

. abstract data will be available for display;

. the conversational language will provide appropriate levels for both the experienced and inexperienced user;

. a paging or browsing facility will enable a user to browse through the files;

. an on-line selective dissemination of information (SDI) service will be provided.

User's interest profiles are stored in the computer. New data entered into the system can be matched against his interest profile and, if appropriate, the details of the document can be stored in the users SDI file. On interrogating his SDI file the user would be informed of the recent additions to his file.

THE FUTURE

As the new and improved services from DSIS come more to the attention of more people, we can expect an increased number of users along with demands for better terminal facilities.

Terminals such as:

(a) high-speed printers

(b) display screens

(c) remote computer devices

– C.P.U.,

– magnetic tapes,

– card or paper-tape punches,

will require faster communication channels than those proposed. Speeds of at least 2000 – 2400 bits per second and higher will be required.

It would be expensive to provide individual communication channels from the central computer to each user. Instead, multiplexing devices enabling a number of lower-speed communication channels to share a single high-speed channel could be provided through local computers already at most establishments.

Figure 6 shows one possible system where numbers of remote users on low-speed lines are connected via a store-and-forward concentrator (SFC) to a high-speed line to the central computing facility.

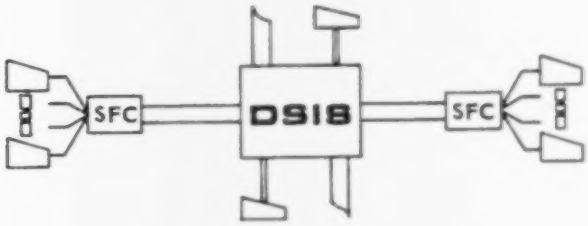


Fig. 6. Future Network.

The SFC, which may be a small computer, will simultaneously collect messages from the user terminals and then transmit them at full speed to the central computer. Outgoing messages will be sent to the SFC, which will distribute the messages to the appropriate user terminals.

The location of such devices and the speeds of the associated communication channels will depend on:

(a) the number of users,

(b) geographical location of users,

(c) type of terminal facilities and their utilisation,

(d) communication channel speeds, availability and costs.

CONCLUSION

Beyond this, the next step will be direct links with other compatible information analysis centres.

Users may be directly connected to several systems or the central computing facility may act as a message switching system. Users may then interrogate other data bases or use their specialized facilities.

Communication and terminal costs are expected to be considerably lower in the near future. All estimates in the present paper were based on current charges of common carriers.

Defence Research and Development Canada | Recherche et développement pour la défense Canada

Canada

Francais | Contact Us | Help | Search | Canada Site

Home | Defence R&D Canada | Defence Site

Browse Indexes

Titles

Authors

Descriptors

Corporate Authors

Corporate Acronyms

Report Numbers

Search Database

Summary

Prior Top **Document:** 3 of 3

Prior Top Bottom Next

Order

Click here to Order

Click here to Download

TITLE:(U) PROCEEDINGS OF SESSION ON INFORMATION RETRIEVAL AT THE 1968 DEFENCE RESEARCH BOARD SYMPOSIUM

PERSONAL AUTHOR: Amey, G.X.

CORPORATE AUTHOR: Defence Research Board, Ottawa ONT (CAN);Defence Research Board, Ottawa ONT (CAN)

ABSTRACT: (U) The following papers are included in the Proceedings: (1) Computer incompatibilities - a software solution (2 Experiences in information retrieval using chemical abstracts, (3) SOCRATES -an information system for the Canadian defence community, (4)File design for Project SOCRATES, (5) Information retrieval for the scientist - a DRES experiment, (6)Communication facilities for an on-line information retrieval system.

DESCRIPTORS: (U) Information retrieval;Documentation;Computer systems programs;Information systems;Data transmission;Socrates (project)

REPORT NUMBER: DR-196;	REPORT DATE: 15 Jun 1969	NUMBER OF PAGES : 46
DRDKIM N°: 69-07728	CANDIS N°: 75363	Format(s):Hardcopy,Originator's fiche received by DSIS;Document Image stored on Optical Disk

Date modified: 2003-04-17

It is recommended you search "Authors" and "Corporate Authors" on their respective "Browse" pages.

Click here if you are having problems displaying text (ie. text is too big, or too small) or page appears to be jumbled.

Important Notices